

Exercise 1 (10p)

Suppose that the probability of five events are

$$P(\text{first}) = 0.3, P(\text{second}) = 0.25 \text{ and}$$

$$P(\text{third}) = P(\text{fourth}) = P(\text{fifth}) = 0.15.$$

Calculate the entropy.

Write down in words what this means.

Exercise 2 (10p)

The data in the table below shows the relationship between the body height and the gender of a group of persons (the records have been sorted with respect to the value of height in cm):

Height	160	165	170	175	180	185
Gender	<i>F</i>	<i>F</i>	<i>M</i>	<i>M</i>	<i>F</i>	<i>M</i>

Calculate the information gain for the potential splitting thresholds (recall that cut points must always lie at class boundaries) and determine the best one.

Exercise 3 (20p)**Canteen - Data.**

A student select a food in canteen as follows (the same data as in Exercise 2 from Homework 2):

Hunger,H	Diet,D	Time,H	Food,F
<i>Great</i>	<i>No</i>	<i>Early</i>	<i>Chips</i>
<i>Great</i>	<i>Yes</i>	<i>Early</i>	<i>Salad</i>
<i>Middle</i>	<i>No</i>	<i>Early</i>	<i>Soup</i>
<i>Middle</i>	<i>Yes</i>	<i>Early</i>	<i>Salad</i>
<i>Great</i>	<i>No</i>	<i>Late</i>	<i>Soup</i>
<i>Great</i>	<i>Yes</i>	<i>Late</i>	<i>Salad</i>
<i>Small</i>	<i>Yes</i>	<i>Late</i>	<i>Salad</i>
<i>Small</i>	<i>No</i>	<i>Late</i>	<i>Salad</i>
<i>Small</i>	<i>No</i>	<i>Early</i>	<i>Salad</i>
<i>Middle</i>	<i>Yes</i>	<i>Late</i>	<i>Salad</i>

Predict with ID3 algorithm the following cases:

- a) H = Great, D = No, Time = Early
- b) H= Middle, D= Yes, Time = Late

Exercise 4 (20p)

Consider the artificial data set shown in following table (the same data as in Exercise 5 from Homework 2)

Time with Emp.,T	Size of Loan,S	Home owner,H	Default,D
5	10.000	1	0
20	10.000	1	0
1	25.000	1	0
1	15.000	3	0
15	2.000	2	0
6	12.000	1	0
1	5.000	2	1
12	8.000	2	1
3	10.000	1	1
1	5.000	3	1

The variables in the columns 1 to 3 are:

- time with the current employer, T , in years;
- size of loan requested, S , in dollars;
- H , where the applicant is a home-owner(1), rental tenant (2), or other (3)

- Find a tree, which will allow prediction of variable D for future customers, where D is default on a bank loan (the last column, labelled 1 for default and 0 for non-default)
- Determine a relative error rate.
- Prune a tree with cost complexity parameter.
- Suppose the new application form is received from an applicant who has been with his or her employer 8 years, is seeking a loan of 10.000, and is a home-owner ($H = 1$) How will you classify this customer?

Exercise 5 (20p)

For the **Pen-Based Recognition of Handwritten Digits Data Set** in the UCI repository construct the decision tree so that it does classification of handwritten digits.

Exercise 6 (20p)

Generate the dataset (as in Exercise 6 in the Homework 1) with 50 data for each of the following two classes:

$$X_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

$$X_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),$$

with

$$\boldsymbol{\mu}_1 = \begin{pmatrix} -3 \\ 7 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 7 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 7 & 5 \\ 5 & 7 \end{pmatrix}$$

- a) Apply CART classification.
- b) Analyse the result.